Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise sentimental textual de tweets em língua portuguesa sobre a pandemia de covid-19 no Brasil

Artur F. S. Zorron

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinícius Pereira Gonçalves

Coorientador
Prof. Dr. Geraldo Pereira Rocha Filho

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise sentimental textual de tweets em língua portuguesa sobre a pandemia de covid-19 no Brasil

Artur F. S. Zorron

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinícius Pereira Gonçalves (Orientador)
ENE/UnB

Prof. Dr. Vinícius Pereira Gonçalves
ENE/UnB

Prof. Dr. Geraldo Pereira Rocha Filho
CIC/UnB

Prof. Dr. Mariana Cabral Falqueiro
ENE/UnB

Prof. Dr. Marcelo Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 23 de fevereiro de 2023

Dedicatória

Gostaria de dedicar este trabalho a todas as pessoas que durante a pandemia de covid-19 no Brasil sofreram de alguma forma, seja pela perda de algum ente querido, pelo sentimento de injustiça e impotência com todas as notícias que eram publicadas ou pela própria percepção que esse período poderia ter sido muito menos caótico e conturbado se as autoridades responsáveis no combate a pandemia não tivessem sido omissas ou agido de forma negligente em diversos momentos durante esse período.

Dedico este trabalho a dias melhores!

Agradecimentos

Agradeço primeiro à minha família, meus pais Roberto e Maria Alice e meu irmão Pedro, pela convivência e pelos ensinamentos. Sem o suporte de vocês não teria chegado onde cheguei. Obrigado por acreditarem em mim.

Agradeço também aos amigos da quadra 102 do Sudoeste, minha segunda família, que me acompanham desde os meus primeiros anos de vida até os dias de hoje. Nossa amizade apenas nos fortalece e sou muito grato por todas nossas histórias e nossos momentos.

Agradeço também à monitoria de Algoritmo e Programação de Computadores (APC) a qual fiz parte durante mais de 3 anos e aprendi muito com vários dos melhores programadores que conheço. Foi um honra ter aprendido e ensinado com vocês!

Agradeço também a Empresa Júnior de Computação da UnB (CJR) por me apresentar o universo do Desenvolvimento Web e me mostrar o caminho que sigo hoje, dentro do mercado de trabalho, como Engenheiro de Software. Fiz dentro dela as maiores amizades na UnB dentro da Computação como o grupo Flor do ENEJ, o qual está o meu ciclo mais próximo de amigos da UnB, que me deram todo o suporte para realizar esta graduação assim como este TCC. Obrigado pelo suporte e pelos momentos de descontração tão necessários no Vale da Lua!

Agradeço também a minha namorada, Mariana Jubé, por sempre me incentivar a produzir este trabalho cada vez mais refinado e por sempre me dar todo o suporte para seguir em frente. E também agradeço à sua tia, Cristina Jubé, pela revisão do texto deste trabalho. Obrigado por toda ajuda!

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Este trabalho apresenta uma análise sentimental dos tweets publicados em língua portuguesa, dentro do território brasileiro, durante o período da pandemia de covid-19. O intuito deste trabalho é analisar como os brasileiros se sentiram emocionalmente quando publicavam textos mencionando a pandemia de covid no Twitter. Os resultados deste trabalho vieram dos resultados do modelo ROBERTA, um analisador sentimental textual. Este trabalho categoriza os textos como positivos, negativos ou neutros e analisa suas proporções quantitativamente e qualitativamente. Para trabalhos futuros, pode-se categorizar a mesma base de textos utilizando conjuntos maiores e mais complexos de emoções, como tristeza, felicidade, raiva, surpresa, etc.

Palavras-chave: análise sentimental, twitter, covid, brasil, nlp

Abstract

This work presents a sentimental analysis of tweets published in Portuguese, within the Brazilian territory, during the period of the covid-19 pandemic. The purpose of this work is to analyze how Brazilians felt emotionally when they published texts mentioning the covid pandemic on Twitter. The results of this work came from the results of the ROBERTA model, a sentimental textual analyzer. This work categorizes texts as positive, negative or neutral and analyzes their proportions quantitatively and qualitatively. For future work, the same base of texts can be categorized using larger and more complex sets of emotions, such as sadness, happiness, anger, surprise, etc.

Keywords: sentimental analysis, twitter, covid, brazil, nlp

Sumário

Lista de Figuras

Lista de Abreviaturas e Siglas

API Interface de Programação de Aplicação.

BERT Representações de Codificador Bidirecional de Transformers.

IA Inteligência Artificial.

NLP Processamento de Linguagem Natural.

ROBERTA Abordagem BERT Otimizada Robustamente.

Capítulo 1

Introdução

Com o início da pandemia de covid-19 no Brasil, foi estabelecida, pelas autoridades, a medida de isolamento social, proposta utilizada para frear a disseminação do vírus da covid-19. Com menos contato pessoal, comunicar-se pela internet e pelas redes sociais tornou-se algo cada vez mais comum. Uma das redes sociais mais utilizadas nesse período foi o Twitter, rede social baseada em postagem de mensagens de texto curtas para consumo rápido. Esses textos são chamados de tweets.

O objetivo deste trabalho é analisar como os brasileiros se sentiram durante o período de pandemia no Brasil utilizando, como base, tweets escritos em português, publicados no período de 2020 a 2022, e que abordam o tema da pandemia. O intuito é rotular esses tweets em positivo, negativo ou neutro utilizando uma inteligência artificial.

A hipótese inicial deste trabalho é que a análise sentimental textual feita pela inteligência artificial, na base de dados contendo tweets publicados no Brasil durante a pandemia de covid-19, seja majoritariamente negativa. Essa hipótese inicial foi criada a partir de alguns fatores. O primeiro fator é a própria experiência de passar por uma pandemia que causou afastamento de pessoas, internações e até mortes, tanto no Brasil quanto no mundo. O segundo fator, mais específico da situação do Brasil, foi a gestão da pandemia pelo governo federal a qual foi muito criticada pela imprensa nacional e internacional, pela demora no reconhecimento da pandemia e de sua gravidade, pelo negacionismo científico no combate à pandemia, pelo atraso da compra de vacinas, ou pelo desprezo de autoridades responsáveis pela gestão do combate à pandemia pelos sentimentos de perda e de luto da população.

Alguns trabalhos importantes para a produção deste artigo foram Mansoor et al. [1] e Kausar et al. [2]. O primeiro trabalho aborda o tema da análise sentimental global de tweets na língua inglesa sobre a pandemia, por isso foi muito útil para analisar graficamente os sentimentos rotulados a partir dos textos dos tweets. Já o segundo trabalho trouxe uma proposta parecida com a do primeiro, porém foi uma análise sentimental de

tweets do início da pandemia e com uma rotulação não binária de sentimentos, como raiva, surpresa, desgosto, etc.

Para este trabalho, utilizou-se uma API do Twitter para coleta de tweets em português sobre a pandemia no Brasil para criação da base de dados. Com a base de dados criada, pré-processou-se os tweets e os traduziu para a língua inglesa. Após a tradução, aplicou-se algoritmos de rotulação textual em análise sentimental nos tweets dessa base utilizando o modelo de inteligência artificial ROBERTA. Após a rotulação dos textos, os resultados coletados revelaram que a população, de forma majoritária, teve percepções negativas e neutras acerca da pandemia, com menor grau de avaliações positivas sobre esse período em nosso país.

Este trabalho foi dividido em cinco capítulos. No primeiro capítulo, são abordados, de maneira generalista, diversos aspectos do trabalho. No segundo capítulo, é apresentado o referencial teórico-metodológico, o qual aborda os conceitos principais deste artigo além dos trabalhos que serviram de base e inspiração para a produção deste. No terceiro capítulo, expõe-se a proposta do modelo utilizado neste trabalho, assim como a arquitetura criada e as ferramentas que auxiliaram no processo de desenvolvimento do código-fonte desta pesquisa. No quarto capítulo, são divulgados os resultados experimentais e as métricas de avaliação do trabalho. No quinto capítulo, são apresentadas as impressões finais acerca do tema e ainda possíveis continuidades dentro do trabalho proposto.

Capítulo 2

Referencial teórico-metodológico

Neste capítulo, serão abordadas algumas questões referenciais a fim de trazer maior embasamento para o leitor ou para a leitora deste trabalho, para que possam, assim, compreender melhor quais os pontos teóricos mais importantes e também quais outros estudos influenciaram na produção desta pesquisa.

2.1 Conceitos introdutórios

Como este trabalho foi construído a partir do consumo de diversas teorias diferentes e de diversos trabalhos relacionados, nesta seção serão explicados de maneira breve, os conceitos de Interface de Programação de Aplicação (API), Processamento de Linguagem Natural (NLP) e análise sentimental.

2.1.1 API

Segundo a própria AWS [3], APIs são mecanismos que permitem que dois componentes de software se comuniquem usando um conjunto de definições e protocolos.

Logo, a API é uma ferramenta de comunicação entre sistemas distintos para compartilhamento de dados. Uma API utiliza de requisições HTTP como GET, POST, PUT e DELETE para interferir em um sistema de maneira externa e objetiva.

Neste trabalho, utilizou-se a API do Twitter para consumir tweets escritos em português durante a pandemia no Brasil, a fim de popular a base de dados utilizada na análise textual.

Um livro muito explicativo sobre a estrutura e funcionamento de APIs é o *REST API Design Rulebook* [4] da Editora O'REILLY.

2.1.2 NLP

De acordo com a IBM [5], NLP refere-se ao ramo da Inteligência Artificial (IA), dentro da ciência da computação, preocupado em fornecer aos computadores a capacidade de entender textos e palavras faladas da mesma forma que os seres humanos.

Desse modo, NLP é um conceito utilizado para extrair informações de textos de uma maneira analítica e gerar entendimento ao leitor com o objetivo de se aproximar, ou até ultrapassar, de uma explicação humana.

Neste trabalho, utilizou-se NLP na análise textual dos tweets da base de dados a fim de fornecer uma análise sentimental acerca dos textos coletados.

Um material muito interessante para o estudo do processamento de linguagem natural é um outro livro da Editora O'REILLY chamado *Practical Natural Language Processing* [6].

2.1.3 Análise sentimental

Pela definição da NVIDIA [7], a análise de sentimento é a interpretação e classificação automatizada de emoções (geralmente positivas, negativas ou neutras) a partir de dados textuais, como críticas escritas e postagens em mídias sociais.

Assim, análise sentimental é uma interpretação, nesse caso computacional, de informações textuais. Ela tem o intuito de rotular textos com certas emoções, se aproximando do entendimento humano de leitura ou interpretação.

Neste trabalho, utilizou-se análise sentimental para rotular os textos dos tweets coletados em positivo, negativo ou neutro, a fim de gerar métricas emocionais da população durante a pandemia.

Uma indicação de estudo para quem quer entender um pouco mais sobre análise sentimental aplicada na redes sociais é o *Sentiment Analysis in Social Networks* [8], também da Editora O'REILLY.

2.2 Trabalhos relacionados

Nesta seção, serão apresentados alguns trabalhos que foram importantes para a escrita deste artigo, suas contribuições e alterações, com elogios e críticas para o refinamento desta pesquisa.

Mansoor et al. [1], em *Global sentiment analysis of COVID-19 tweets over time* retratam a percepção sentimental mundial via tweets na língua inglesa sobre a pandemia de covid-19. Trazem uma relação temporal entre número de infectados com porcentagem de tweets de cunho negativo nos Estados Unidos, na Índia e no Brasil. Ademais levantam

pontos de análise sobre as porcentagens periódicas de tweets negativos de cada país com acontecimentos locais de cada país. Um exemplo é o aumento de tweets positivos em abril, mesmo mês em que o governo federal disponibilizou o auxílio emergencial distribuído durante a pandemia no Brasil. Todavia, Mansoor et al. não trazem uma análise focada em tweets em língua portuguesa publicados por perfis brasileiros.

Já Kausar et al. [2], em *Public sentiment analysis on Twitter data during COVID-19 outbreak*, produzem um estudo de análise sentimental de tweets publicados no início da pandemia, de janeiro a junho de 2020. Esse trabalho traz consigo o conceito de nuvens de palavras, apresentando diferentes análises de palavras mais utilizadas em determinadas buscas, porém muitas das palavras acabam sendo muito mais relevantes em sua regionalidade, como a palavra *Trump*, referente ao ex-presidente norte-americano Donald Trump, que não se aplica ao cenário brasileiro de publicação de tweets. Um equivalente a essa palavra muito utilizada foi *Bolsonaro*, dirigindo-se ao ex-presidente do Brasil, Jair Bolsonaro, o qual era presidente exatamente no período mais crítico da pandemia de covid-19.

Outro trabalho interessante foi o de Melo et al. [9], *Comparing news articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach*, que comparou análises sentimentais provenientes do Twitter com artigos de notícias sobre a pandemia no Brasil. O diferencial desse trabalho é o foco da análise sentimental voltada para o Brasil, porém esse trabalho não comparou os diferentes meios de comunicação. Mesmo com focos diferentes, o trabalho de de Melo et al. auxiliou bastante este artigo com suas nuvens de palavras, identificando palavras-chave nas publicações de tweets referentes à pandemia no Brasil.

Além de trabalhos acerca do Brasil ou Estados Unidos, *Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets*, de Sitaula et al. [10] aborda diferentes formas de analisar sentimentos alterando o método de *deep learning* implementado para tal, sobre tweets publicados no Nepal. Isso contribui para a compreensão das diferentes percepções sobre a pandemia ao alterar o país estudado das divergências em relação ao país, cerne da pesquisa, e ao uso de diferentes métodos de *deep learning*.

Todos os trabalhos citados acima analisaram a pandemia em seus respectivos países de forma mais generalista, não se aprofundando em um tema específico. O estudo que influenciou a escolha da linha de pesquisa deste artigo foi *Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis* do Liu et al. [11]. Nesse estudo, é analisado como a sociedade se portou emocionalmente sobre o tema da vacinação no mundo. Adentrar neste tema ao analisar tweets sobre a covid-19 é de extrema importância, já que essa foi a primeira pandemia global na história em que a humanidade tinha conhecimento do que são os vírus e de como combatê-los de maneira eficaz, com vacinas e isolamento social. Esse estudo trouxe análise sobre o sentimento da população sobre as

vacinas e isso auxiliou entender de forma geral se minha hipótese incial faria sentido, ao predizer que a avaliação seria mais negativa que positiva, assim como o trabalho de Liu et al..

Capítulo 3

Análise sentimental textual de tweets na língua portuguesa sobre a pandemia de covid-19 no Brasil

Neste capítulo, serão apresentados a arquitetura principal deste artigo, quais os sistemas relacionados dentro deste trabalho e como se relacionam, além de explicar melhor cada uma das etapas das análises dos dados.

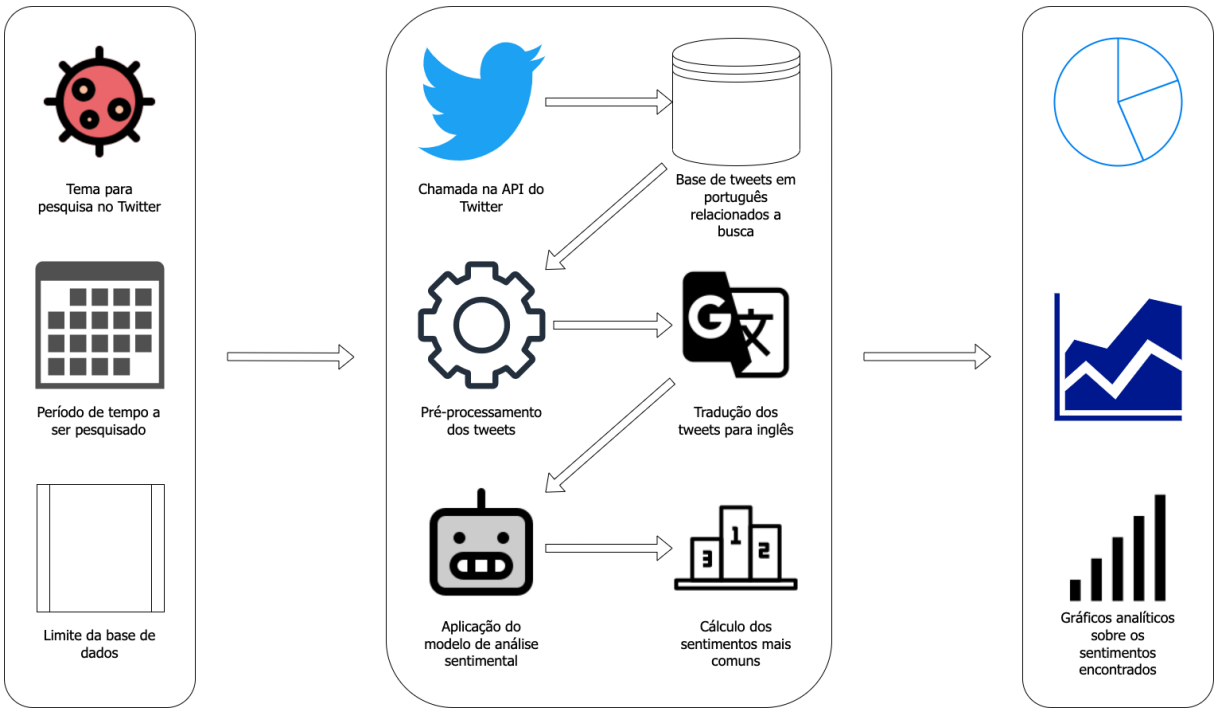

Figura 3.1: Arquitetura do programa e seus componentes

Na figura 3.1, observam-se as três etapas principais da arquitetura do trabalho. Inicia-se com o tema a ser pesquisado na API do Twitter, o período que os Tweets foram

publicados e o limite de tweets a serem adicionados à base de dados. Após definir as métricas de busca dos tweets, aplicamos a segunda etapa do trabalho, que consiste de seis componentes principais. O primeiro é a chamada na API do Twitter com os parâmetros da primeira etapa. O segundo é o armazenamento de todos os tweets coletados. O terceiro é o pré-processamento da base de tweets. Já o quarto é a tradução dos tweets para a língua inglesa. O quinto componente é a aplicação do modelo de análise sentimental. E, por último, o sexto calcula as probabilidades e porcentagens dos tweets terem sentimentos específicos. Ao fim da segunda etapa, a terceira é análise gráfica dos resultados do último componente da etapa anterior.

As duas primeiras etapas da arquitetura serão exploradas neste capítulo, e os resultados, no Capítulo 4.

3.1 Coleta de tweets para base de dados

Este trabalho consiste em uma análise sentimental de tweets em português no Brasil, durante o período da pandemia. A primeira parte da pesquisa foi coletar esses tweets da base do Twitter.

Para isso, utilizou-se uma biblioteca chamada *snscrape* para realizar a comunicação via API com o Twitter. Ela funciona passando uma query à chamada do método e este retorna com uma lista de tweets que se encaixam nessa query. A query é uma string com algumas tags demonstrando informações necessárias no tweet para ele entrar na base de dados, atuando como um filtro sobre a base de tweets do Twitter.

```
import snscrape.modules.twitter as sntwitter
query = 'covid OR pandemia lang:pt until:2023-01-01 since:2020-01-01'
limit = 100000
dates = []
locations = []
tweets_pt = []
for tweet in sntwitter.TwitterSearchScraper(query).get_items():
    if len(tweets_pt) >= limit:
        break
    if tweet.user.location.find('Brasil') != -1:
        dates.append(tweet.date)
        locations.append(tweet.user.location)
        tweets_pt.append(tweet.rawContent)
```

Figura 3.2: Coleta dos tweets via API e armazenamento das datas, localizações e textos

Nesse trecho de código, buscam-se por tweets cujo texto possua as palavras covid ou pandemia, que seja escrito em português e tenha sido postado de 2020 a 2022. Com isso, determina-se um limite de 100 mil tweets a serem coletados que se encaixem nesses pré-requisitos. Utilizando a biblioteca *snscrape*, popula-se a base de dados com tweets vindos do filtro.

3.2 Pré-processamento de tweets

Ao coletar os tweets da base do Twitter, vêm com eles várias informações sensíveis aos usuários, como *usernames* e *links* dentro das mensagens. Para resolver isso, realiza-se um pré-processamento nesses textos a fim de mascarar essas informações, visto que elas não são importantes dentro das análises.

```
tweets_proc = []
for tweet in tweets_pt:
    tweet_words = []
    for word in tweet.split():
        if word.startswith('@') and len(word) > 1:
            word = '@user'
        elif word.startswith('http'):
            word = 'http'
        tweet_words.append(word)
    tweets_proc.append(" ".join(tweet_words))
```

Figura 3.3: Pré-processamento dos tweets removendo menções e link

Nesse trecho de código, itera-se por todos os tweets com menções a usuários ou urls. Caso encontre uma menção a um usuário, deve-se alterar para *@user*. Caso seja encontrada uma url iniciada com http, altera-se toda a url para apenas *http*.

3.3 Tradução de tweets de português para inglês

Para analisar os tweets sentimentalmente, é preciso adequá-los aos moldes que o modelo analítico pede. Neste artigo, utiliza-se uma variante do modelo Representações de Codificador Bidirecional de Transformers (BERT), chamada de Abordagem BERT Otimizada Robustamente (ROBERTA). Esse modelo consegue realizar analíses sentimentais binárias de valência probabilísticas na língua inglesa. Todavia, os tweets coletados na base de dados para essa pesquisa são todos em português. Portanto, devem-se traduzir esses tweets utilizando uma biblioteca de tradução chamada *deep_translator*.

```
from deep_translator import GoogleTranslator
tweets_en = GoogleTranslator('pt', 'en').translate_batch(tweets_proc)
```

Figura 3.4: Tradução dos tweets com script em *batch*

No trecho de código acima, faz-se a tradução da lista de tweets pré-processados utilizando um script em batch do método *GoogleTranslator()* da biblioteca *deep_translator*.

3.4 Análise sentimental de tweets

Já com os tweets pré-processados e traduzidos, pode-se agora aplicar um modelo de análise sentimental sobre a base de tweets para analisar a probabilidade do tweet ser positivo, negativo e neutro.

```
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
roberta = 'cardiffnlp/twitter-roberta-base-sentiment-latest'
tokenizer = AutoTokenizer.from_pretrained(roberta)
model = AutoModelForSequenceClassification.from_pretrained(roberta)
labels = ['Negative', 'Neutral', 'Positive']
```

Figura 3.5: Criação do modelo, do tokenizador e das categorias sentimentais

Nesse trecho de código, utiliza-se o modelo ROBERTA para criar o modelo e o tokenizador, além das categorias negativo, neutro e positivo.

```
from scipy.special import softmax
scores = []
for tweet in tweets_en:
    encoded_tweet = tokenizer(tweet, return_tensors='pt')
    output = model(**encoded_tweet)
    output_score = output[0][0].detach().numpy()
    scores.append(softmax(output_score))
```

Figura 3.6: Cálculo das probabilidades sentimentais de cada tweet

No trecho de código acima, calcula-se a probabilidade de cada tweet estar relacionado a um dos sentimentos possíveis.

Com os resultados já coletados, é feito apenas um tratamento nos dados para agregar os resultados a seus respectivos tweets.

3.5 Criação da base de dados final

Após coletar todos dados das etapas anteriores, é necessário agregar todas essas informações em um só lugar para gerar gráficos a fim de analisar visualmente os resultados obtidos.

```
import pandas as pd
df = pd.DataFrame({
    'date': dates,
    'location': locations,
    'tweets_pt': tweets_proc,
    'tweets_en': tweets_en,
    'negative': [score[0] for score in scores],
    'neutral': [score[1] for score in scores],
    'positive': [score[2] for score in scores],
    'label': [labels[score.argmax()] for score in scores]
})
df.to_csv('database.csv')
```

Figura 3.7: Cálculo das probabilidades sentimentais de cada tweet

Neste trecho de código acima, cria-se um dataframe utilizando a biblioteca *pandas* adicionando, para de cada tweet, as colunas data, localização, texto em português, tradução em inglês, probabilidade negativa, probabilidade neutra, probabilidade positiva e a categoria que possuiu a maior probabilidade. Com essas informações, pode-se criar gráficos para enfim analisar como os brasileiros se sentiram durante esse período da pandemia.

Capítulo 4

Resultados experimentais

4.1 Experimentação

O planejamento deste trabalho foi utilizar a rede social Twitter como base de dados; coletar tweets em português, publicados no Brasil durante o período da covid-19, que tenham relação com essa temática; pré-processar essa base de tweets para remover menções a usuários e links publicados, a fim de manter o sigilo e a segurança de quem publicou; traduzir os textos desses tweets para a língua inglesa; aplicar um modelo probabilístico de análise sentimental sobre cada tweet, rotulando em porcentagem a probabilidade de ele ser negativo, neutro ou positivo, gerar uma base de dados compilada com todas essas informações a fim de produzir gráficos analíticos sobre a pesquisa; e, ao final, apresentar o resultado como médias das probabilidades de cada rótulo sentimental.

O algoritmo foi executado inúmeras vezes a fim de testar diferentes cenários e entradas para o consumo da API do twitter, para a rotulação dos tweets e para os resultados finais. Foram alterados a quantidade de tweets a serem coletados, quais informações seriam coletadas dos tweets, o conteúdo dos textos publicados, os períodos de publicação, os locais onde o usuário publicou, a língua do texto publicado e qual modelo analítico utilizar para rotulação da base de dados.

4.2 Resultados

A partir do planejamento desses experimentos, chegaram-se a alguns resultados importantes. Primeiro, era necessário avaliar a acurácia do modelo analítico ROBERTA. Para isso, foi criado um formulário aberto às pessoas para coletar suas percepções sentimentais sobre 20 tweets escolhidos de dentro da base de dados a fim de comparar com os resultados da ROBERTA.

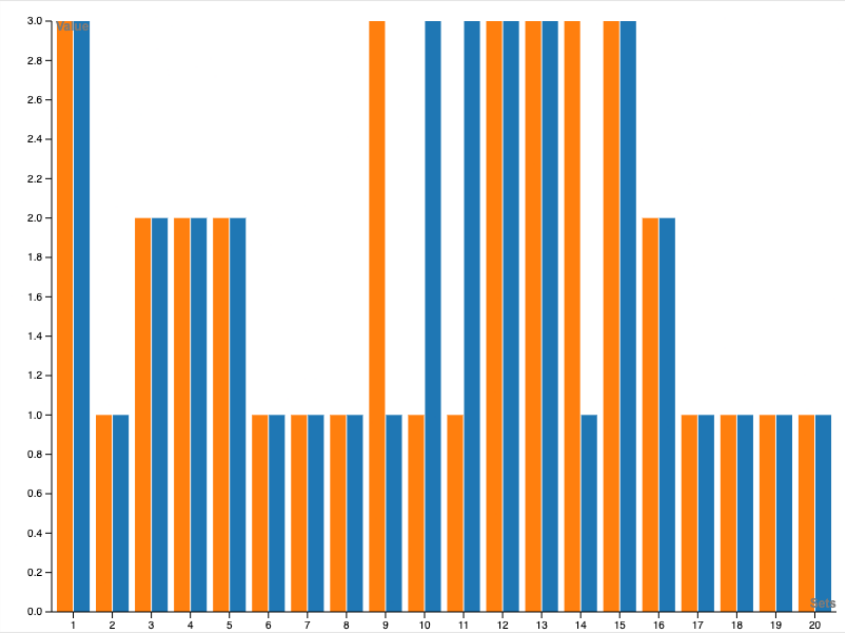

Figura 4.1: Comparação entre respostas do formulário e respostas do modelo BERT. Os valores 1, 2 e 3 são, respectivamente, negativo, neutro e positivo. As cores laranja e azul representam, respectivamente, as respostas do formulário e os resultados do modelo ROBERTA.

Neste gráfico, é possível analisar que a acurácia da ROBERTA foi alta, 80% de precisão, tendo um elevado grau de similaridade com as respostas dos participantes. Dentre os vinte tweets, obtiveram o mesmo resultado em dezesseis deles. Nos 4 tweets com respostas distintas, existiram dois cenários. Havia ironia em dois tweets o que pode ter dificultado entender, para o modelo, se era negativo ou positivo. Já em outros dois haviam muitos comentários negativos e no final um comentário positivo, o que pode ter afetado no entendimento do modelo sobre suas definições.

Após comprovar a acurácia do modelo utilizado, podem-se avaliar os resultados provenientes de sua análise sentimental. Uma primeira análise interessante é comparar quantitativamente quantos tweets foram rotulados como negativo, neutro ou positivo.

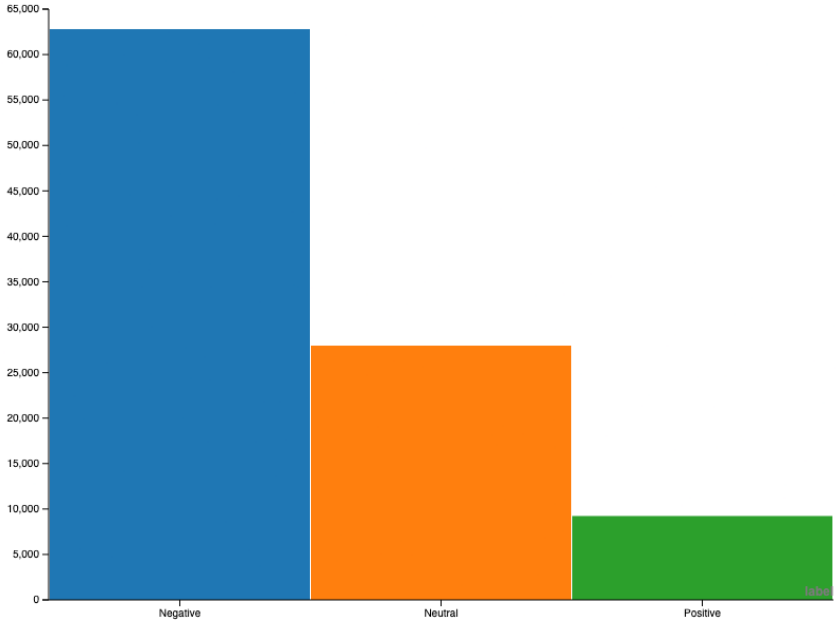

Figura 4.2: Número de tweets de cada sentimento

No gráfico, é possível compreender que os tweets, em maior volume, foram rotulados como negativos, seguidos dos neutros e, por último, os positivos.

É interessante também analisar qualitativamente do que apenas quantitativamente. Para isso, pode-se verificar a média das probabilidades de cada tweet ser um dos rótulos sentimentais. Em grande escala, isto mostra uma análise mais precisa das porcentagens sentimentais da base de dados pois calcula-se a partir de do somatório de cada tweet.

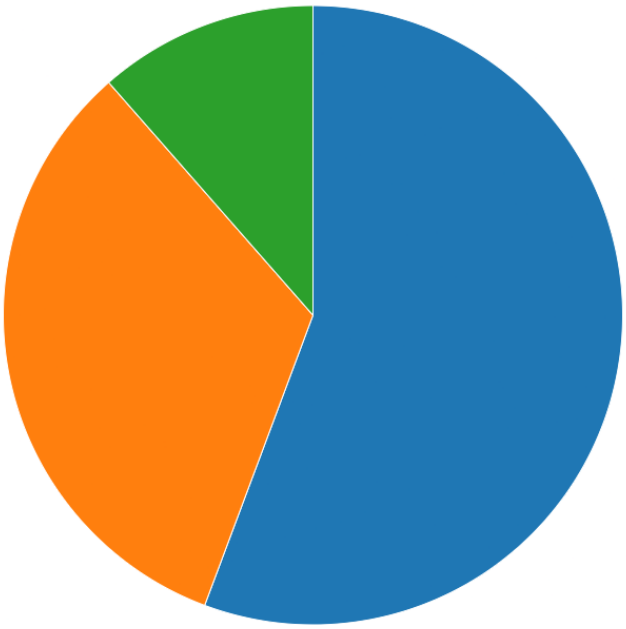

Figura 4.3: Relação percentual do número de tweets de cada sentimento

É possível identificar, no gráfico que os tweets, de maneira percentual, tiveram uma maior chance de serem rotulados como negativos, seguido de neutros e positivos, assim como a análise quantitativa apresentou.

E, para finalizar, é importante analisar quais os períodos que esses tweets foram publicados, pois, apesar de selecionar os anos de 2020 a 2022 (3 anos), o limite de 100 mil tweets pode não ter abrigado esse cálculo por completo. Neste trabalho foram coletados tweets desde o final de dezembro de 2022 até o final de novembro de 2022.

Figura 4.4: Mapa de cores de sentimentos de tweets distribuidos em dias

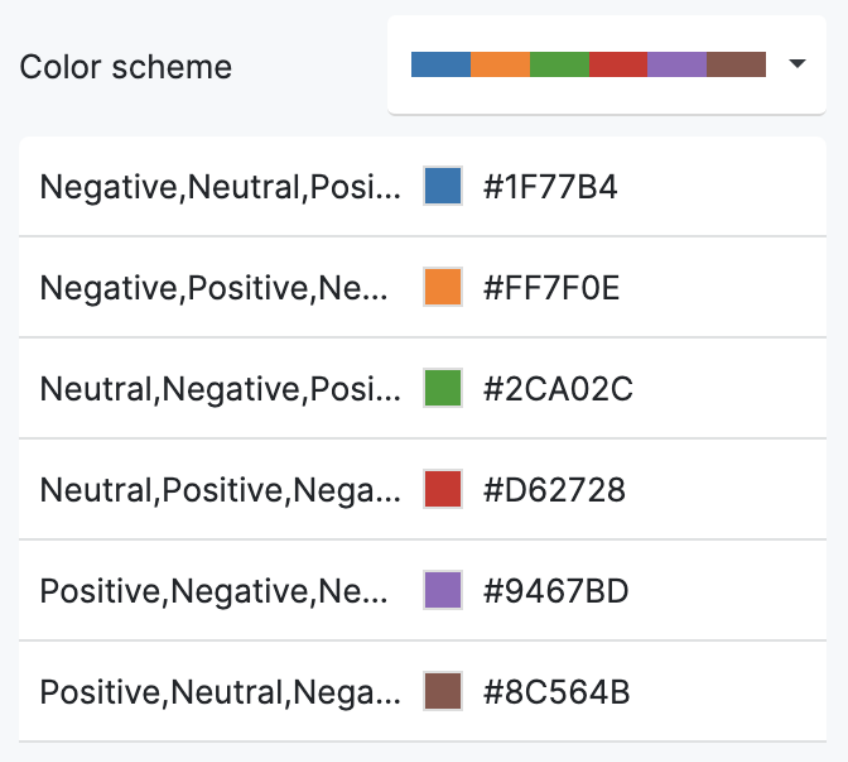

Figura 4.5: Legenda das cores do mapa de cores

Nas imagens acima, é possível analisar que as datas tiveram predominantemente uma quantidade maior de tweets negativos sobre outros tipos de tweets, seguidos de tweets neutros sendo os mais comum em uma data específica.

Capítulo 5

Conclusão

Após a coleta dos resultados experimentais promovida pela análise sentimental sobre a base de dados de tweets, é possível entender e afirmar alguns pontos conclusivos neste trabalho.

É possível identificar um descontentamento da população brasileira quando é mencionado o tema da pandemia de covid-19. Mesmo não coletando todos os tweets relacionados à temática, pois é um problema muito complexo dentro de NLP mapear todos os possíveis tweets relacionados ao tema, já que cada usuário se expressa de maneiras diferentes, utilizando termos diferentes para se dirigir ao tema, pode-se sim afirmar que, no conjunto da obra, houve uma tendência mais negativa do que positiva das expressões da população em geral.

5.1 Trabalhos futuros

Este trabalho não utilizou um modelo de análise sentimental para textos em português. Apesar de existir o modelo BERTIMBAU proposto por Fábio de Souza et al. [12], o qual também faz parte do padrão BERT de modelos, ele não se mostrou preparado para a realização de análises sentimentais para textos na língua portuguesa.

Uma possível continuação deste trabalho seria analisar os tweets da língua portuguesa em um modelo capaz de analisar textos em português, seja utilizando algum modelo já existente ou preparando um para tal.

Outro ponto importante seria analisar temas específicos dentro do contexto da pandemia, como vacinação, gestão da saúde pública, disseminação de fake news, etc. Desse modo, consegue-se compreender resultados ainda mais nichados para cada tema e avaliar melhor os sentimentos expressos pelos brasileiros quando tratam sobre a pandemia no Twitter.

É imprescindível analisar tweets filtrados por datas em que ocorreram marcos importantes durante o período do covid-19 no Brasil, como o dia do primeiro infectado, do primeiro óbito por covid, do atingimento da marca de 100 mil mortos, da crise de oxigênio em Manaus, etc.

Este artigo apresenta análises sentimentais interessantes e pertinentes àqueles que querem analisar melhor os sentimentos dos brasileiros durante o período da pandemia, além de fornecer possíveis caminhos para trabalhos futuros a fim de perpetuar o estudo sobre esse tema que impactou tanto o Brasil.

Referências

[1] Mansoor, Muvazima, Kirthika Gurumurthy, VR Prasad *et al.*: *Global sentiment analysis of covid-19 tweets over time.* arXiv preprint arXiv:2010.14234, 2020. 1, 4

[2] Kausar, Mohammad Abu, Arockiasamy Soosaimanickam e Mohammad Nasar: *Public sentiment analysis on twitter data during covid-19 outbreak.* International Journal of Advanced Computer Science and Applications, 12(2), 2021. 1, 5

[3] AWS. https://aws.amazon.com/what-is/api. 3

[4] O'REILLY. https://www.oreilly.com/library/view/rest-api-design/9781449317904. 3

[5] IBM. https://www.ibm.com/topics/natural-language-processing. 4

[6] O'REILLY. https://www.oreilly.com/library/view/practical-natural-language/9781492054047/. 4

[7] NVIDIA. https://www.nvidia.com/en-us/glossary/data-science/sentiment-analysis. 4

[8] O"REILLY. https://www.oreilly.com/library/view/sentiment-analysis-in/9780128044384. 4

[9] Melo, Tiago de, Carlos MS Figueiredo *et al.*: *Comparing news articles and tweets about covid-19 in brazil: sentiment analysis and topic modeling approach.* JMIR Public Health and Surveillance, 7(2):e24585, 2021. 5

[10] Sitaula, Chiranjibi, Anish Basnet, Ashish Mainali, Tej Bahadur Shahi *et al.*: *Deep learning-based methods for sentiment analysis on nepali covid-19-related tweets.* Computational Intelligence and Neuroscience, 2021, 2021. 5

[11] Liu, Siru e Jialin Liu: *Public attitudes toward covid-19 vaccines on english-language twitter: A sentiment analysis.* Vaccine, 39(39):5499–5505, 2021. 5

[12] Souza, Fábio, Rodrigo Nogueira e Roberto Lotufo: *BERTimbau: pretrained BERT models for Brazilian Portuguese.* Em *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020. 18